LEMBAR KERJA MAHASISWA (LKM)

LK.8 Perancangan Project Data Science

Nama	**: RAHMADANI**
Tanggal	**: 6 Desember 2025**
Kelas	**: 5AI-A**
Judul Project	**: Prediksi Jenis Wilayah Sekolah Menggunakan Random Forest**

A. Instruksi

Peserta diminta untuk merancang sebuah proyek Data Science yang berfokus pada permasalahan di bidang pendidikan. Rancangan proyek ini harus disusun secara sistematis berdasarkan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) yang mencakup enam tahapan utama, yaitu:

1. Business Understanding (Pemahaman Bisnis)
2. Data Understanding (Pemahaman Data)
3. Data Preparation (Persiapan Data)
4. Modeling (Pemodelan)
5. Evaluation (Evaluasi)
6. Deployment (Penerapan)

Pada setiap tahapan, peserta diharapkan dapat:

1. Menjelaskan tujuan dan fokus kegiatan pada tahap tersebut.
2. Menguraikan langkah-langkah yang dilakukan serta teknik atau metode yang digunakan.
3. Menjelaskan jenis dan sumber data yang diperlukan.
4. Menunjukkan hasil atau keluaran yang diharapkan dari tiap tahap.

Gunakan contoh kasus nyata atau permasalahan aktual di dunia pendidikan, seperti: Prediksi prestasi belajar siswa, Analisis tingkat kehadiran, Deteksi dini siswa berisiko tidak lulus, atau Rekomendasi pembelajaran adaptif berbasis data.
Hasil akhir dari tugas ini berupa dokumen rancangan proyek Data Science lengkap yang menggambarkan alur proses dari awal hingga implementasi model, serta menunjukkan bagaimana solusi berbasis data dapat memberikan manfaat nyata bagi peningkatan mutu pendidikan.

B. Format Perancangan

Tahapan CRISP-DM	Instruksi untuk Peserta	Rancangan Implementasi
1. Business Understanding (Pemahaman Bisnis)	1. Pilih konteks pendidikan (contoh: sekolah, universitas, pelatihan). 2. Identifikasi permasalahan yang dapat diselesaikan dengan data science. 3. Rumuskan tujuan bisnis (contoh: meningkatkan prestasi siswa, menurunkan tingkat ketidakhadiran).	Pada project ini konteksnya berada pada dunia pendidikan, khususnya sekolah dasar dan menengah di Indonesia. Data yang digunakan berisi informasi mengenai kondisi sekolah, jumlah peserta didik, jumlah guru, fasilitas, dan indikator kualitas lainnya. Permasalahan yang ingin diselesaikan adalah

		bagaimana memprediksi jenis wilayah sekolah (URBAN atau RURAL) berdasarkan kondisi sekolah tersebut. Tujuan bisnis dari proyek ini yaitu membantu pihak terkait (seperti dinas pendidikan atau sekolah) untuk memahami faktor-faktor apa saja yang paling berpengaruh terhadap kategori wilayah sekolah. Dengan adanya prediksi ini, sekolah di wilayah tertentu bisa dipetakan kebutuhannya, terutama terkait fasilitas pendidikan dan pemerataan kualitas.
2. Data Understanding (Pemahaman Data)	1. Jelaskan sumber data (contoh: data nilai siswa, absensi, data keluarga). 2. Sebutkan jenis data (numerik, kategorikal, teks, waktu). 3. Deskripsikan fitur dan target yang akan digunakan.	Sumber data berasal dari https://data.kemendikdasmen.go.id/dataset/p/asesmen-nasional/rapor-publik-asesmen-nasional-2024-kepala-satuan-pendidikan-2024-indonesia, yaitu dataset resmi yang memuat informasi lengkap mengenai kondisi sekolah di seluruh Indonesia. **Jenis data yang digunakan dalam project ini:** **1. Data Numerik:** jumlah_peserta_didik, jumlah_pendidik, rasio_pendidik_peserta_didik, jumlah_r_kelas, proporsi_pendidik_min_s1, proporsi_pendidik_sertifikasi, jumlah_komp_milik, jumlah_rombel, jumlah_siswa_rombel, jumlah_siswa_penerima_PIP, Rasio PIP, dan semua indikator kualitas seperti: • APC • APK • BBS • BCP • COP • CSV

		• EQC • EQR • HWS • KKG • KPC • KPK • KSA • KSV • LCS • LFO • MIP • NAT • OPC • PBU • PCP • PGP • PIP • PKG • PMU • POT • PPK • PSA • PSV • PPC • RC • RKG • RPI • RSD • SBU • SCO • SKG • SSV • TOC • TOR • Dan seluruh variabel angka lain yang ada di dataset. **2. Data Kategorikal:** jenis_sek, kurikulum, daerah_khusus, wilayah_bagian, jenis_wilayah (target), status_wilayah, sts_sek, ketersediaan_internet, ketersediaan_listrik, pendidikan_sederajat, status sekolah (S/N). kode wilayah: kd_kokab, kd_sekolah, kd_kepsek_an

		Fitur yang digunakan adalah gabungan dari: • Data numerik: jumlah siswa, jumlah guru, rasio, indikator mutu, dan data kuantitatif lain. • Data kategorikal: jenis sekolah, kurikulum, status, daerah khusus, bagian wilayah, dll. • Semua fitur ID tidak digunakan. **Target**: • **jenis_wilayah** → menunjukkan apakah sekolah berada di wilayah Urban atau Rural.
3. Data Preparation (Persiapan Data)	1. Tuliskan langkah pembersihan data: hapus duplikat, tangani nilai kosong, dan outlier. 2. Transformasi data: normalisasi, encoding data kategorikal.	**1. Menghapus data duplikat** Beberapa baris data yang identik dihapus agar tidak memengaruhi hasil training. **2. Menangani nilai kosong** Untuk kolom numerik → diisi dengan median atau mean. Untuk kolom kategorikal → diisi kategori terbanyak (mode). **3. Menangani outlier** Memeriksa fitur jumlah siswa, jumlah guru, rasio, dan fitur-fitur indikator lain untuk memastikan tidak ada nilai ekstrem yang merusak model. **4. Encoding data kategorikal** Mengubah kategori menjadi angka menggunakan: • Label Encoding • One Hot Encoding (jika diperlukan) Contoh yang di-encode: kurikulum, jenis_sek, status sekolah, internet/listrik. **5. Normalisasi (opsional)** Karena Random Forest tidak terlalu sensitif terhadap skala, normalisasi hanya diterapkan jika ada fitur tertentu yang perbedaannya terlalu besar.

4. Modeling (Pemodelan)	1. Pilih algoritma yang sesuai (contoh: Decision Tree, Random Forest, Logistic Regression). 2. Jelaskan alasan pemilihan algoritma.	Algoritma yang digunakan adalah **Random Forest Classifier**, karena cocok untuk dataset dengan banyak fitur numerik dan kategorikal, serta menghasilkan performa yang stabil. **Langkah-langkah pemodelan:** 1. Membagi data menjadi training dan testing (80:20). 2. Melatih model Random Forest menggunakan data training. 3. Mengatur parameter seperti jumlah tree, depth, dan criterion jika diperlukan. 4. Melakukan prediksi pada data testing. 5. Menyimpan model terbaik untuk evaluasi. **Alasan pemilihan Random Forest:** • Tahan terhadap noise dan missing value kecil • Tidak membutuhkan scaling • Bisa menangani fitur numerik dan kategorikal • Menghasilkan feature importance untuk melihat faktor paling berpengaruh • Umumnya akurat di dataset pendidikan
5. Evaluation (Evaluasi)	Pilih metode evaluasi yang akan digunakan misalkan menggunakan cross-validation atau confusion matrix.	Pada tahap evaluasi, performa model diuji menggunakan tiga metode utama, yaitu **Accuracy Score**, **Confusion Matrix**, dan **Classification Report**. **Accuracy Score** digunakan untuk melihat seberapa besar persentase prediksi model yang benar. **Confusion Matrix** digunakan untuk mengetahui distribusi prediksi benar dan salah pada tiap kelas (Urban dan Rural). Sedangkan **Classification Report** memberikan informasi lebih detail melalui nilai **Precision**, **Recall**, dan **F1-Score** sehingga dapat terlihat

		apakah model bekerja seimbang untuk kedua kelas. Hasil evaluasi dari ketiga metode ini membantu memastikan bahwa model Random Forest dapat memprediksi jenis wilayah sekolah secara akurat dan tidak bias ke salah satu kelas.
6. Deployment (Penerapan / Implementasi)	Buat rancangan deploymentnya tampilan interface nya	Model akan di-deploy menggunakan Gradio sehingga dapat digunakan dengan mudah. **Desain Interface Gradio:** **Bagian Input (yang diisi oleh pengguna)** Ini semacam form berisi kolom-kolom yang bisa diisi, misalnya: • jumlah peserta didik • jumlah pendidik • proporsi guru S1 • kurikulum • status sekolah • indikator mutu (misal APK, BBS, COP, dll) Pengguna akan memasukkan angka atau memilih kategori di setiap kolom. **Tombol "Prediksi"** Setelah data diisi, pengguna menekan tombol **Submit / Predict**. **Bagian Output (hasil dari model)** Setelah tombol ditekan, model akan menampilkan: • **Prediksi wilayah sekolah → Urban / Rural** • **Probabilitas prediksi** • **Fitur yang paling berpengaruh** terhadap prediksi (feature importance)